Towards Explainable NLP: A Generative Explanation Framework for Text Classification

Authors:
Hui Liu, Qingyu Yin,
William Yang Wang

Presented by:
Janaki Viswanathan,
Explainability Methods for NNs
Saarland University - [WS 20/21]

Overview

- Motivation
- Proposed solution
- Generative Explanation Framework (GEF)
    - Intuition
    - Base Classifier and Generator
    - Explanation Factor
    - Minimum Risk Training
- Experiment results

Overview

- Motivation
- Proposed solution
- Generative Explanation Framework (GEF)
    - Intuition
    - Base Classifier and Generator
    - Explanation Factor
    - Minimum Risk Training
- Experiment results

Motivation

- Deep learning methods have produced state-of-the-art results in many NLP tasks

Motivation

- Deep learning methods have produced state-of-the-art results in many NLP tasks

But they are blackboxes for human beings!

Motivation

- **Example:** Rating a product (by a human)

Motivation

- **Example:** Rating a product (by a human)
    - Review: "The phone feels sturdy, looks premium, and works great. For a mostly business user like me, these features, plus all the little software enhancements and customizations that I am allowed to do make it a very attractive device......."

Motivation

- **Example:** Rating a product (by a human)
  - Review: “The phone feels sturdy, looks premium, and works great. For a mostly business user like me, these features, plus all the little software enhancements and customizations that I am allowed to do make it a very attractive device.......”

  - Attribute scoring:

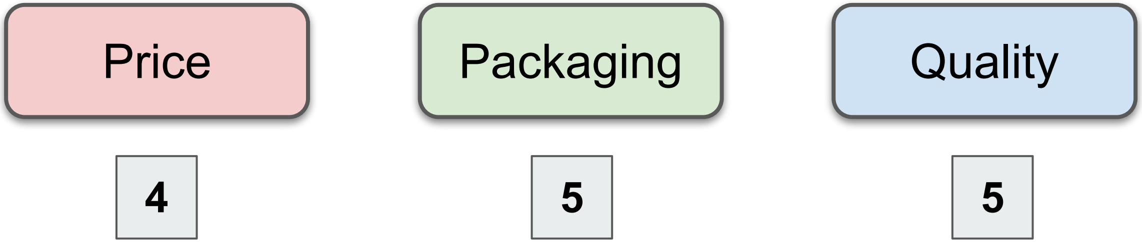
Motivation

- **Example:** Rating a product (by a human)
  - Review: “The phone feels sturdy, looks premium, and works great. For a mostly business user like me, these features, plus all the little software enhancements and customizations that I am allowed to do make it a very attractive device.......”
  - Attribute scoring:

Price	Packaging	Quality
4	5	5

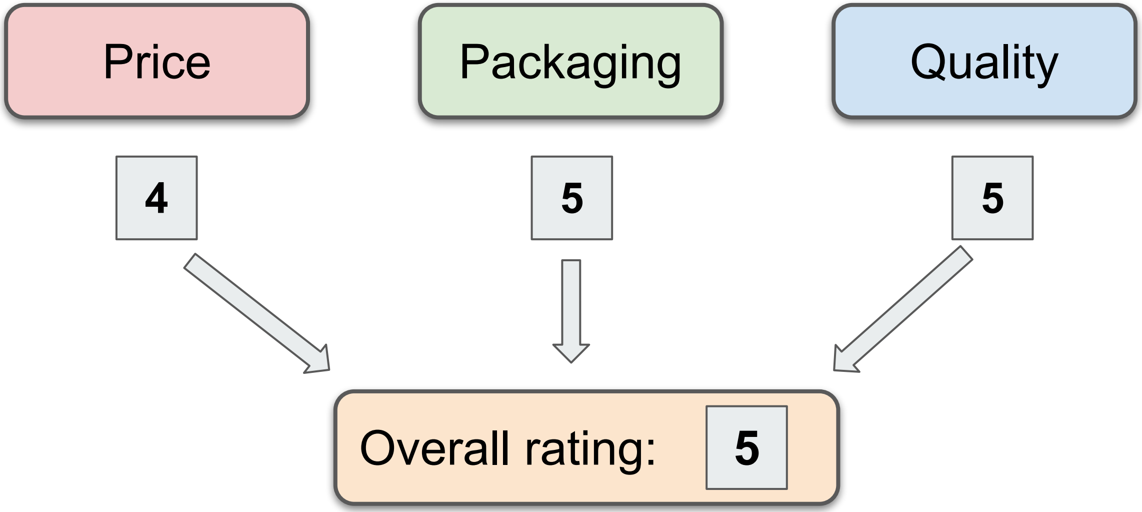
Motivation

- **Example:** Rating a product (by a human)



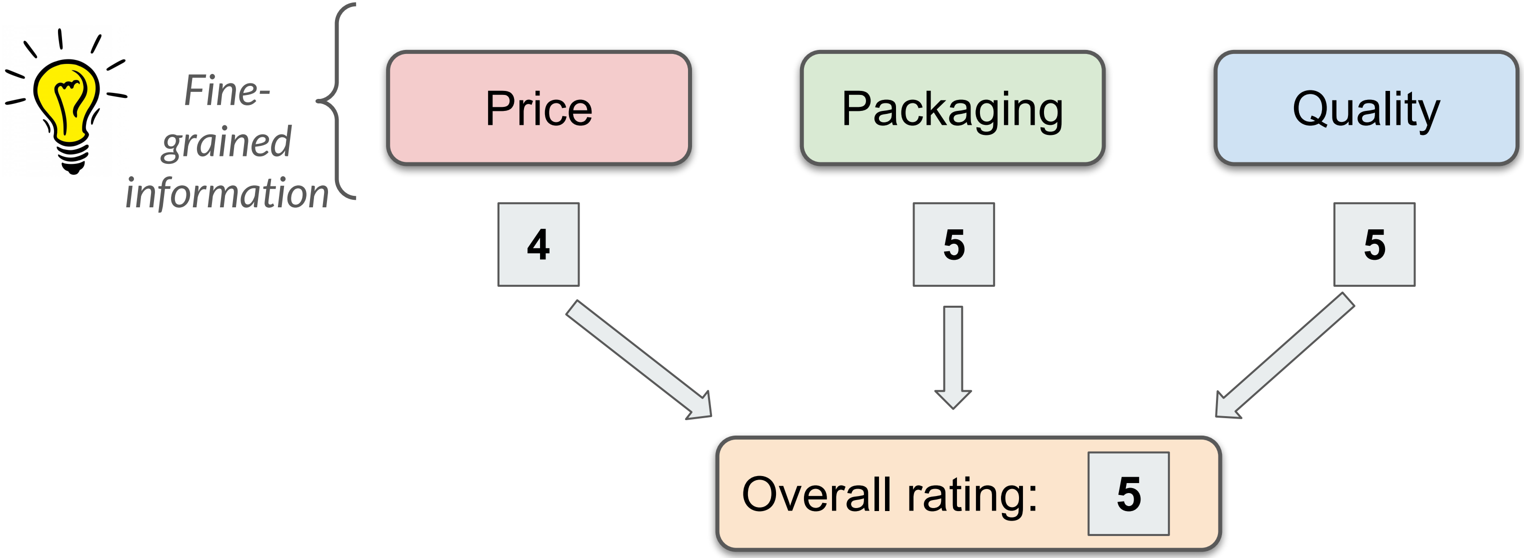
Motivation

- **Example:** Rating a product (by a human)



Motivation

- **Example:** Rating a product (by a human)



Motivation

- Deep learning methods have produced state-of-the-art results in many NLP tasks

Ability to explain rationale is essential for an NLP system!

Overview

- Motivation
- **Proposed solution**
- Generative Explanation Framework (GEF)
  - Intuition
  - Base Classifier and Generator
  - Explanation Factor
  - Minimum Risk Training
- Experiment results

Proposed solution

- **Goal:**

  To build trustworthy **explainable** text classification **model** capable of explicitly **generating fine-grained information** for explaining their predictions

Proposed solution

- Generative Explanation Framework (**GEF**)
  - Makes classification decisions
  - Generates fine-grained explanations

Overview

- Motivation
- Proposed solution
- **Generative Explanation Framework (GEF)**
    - Intuition
    - Base Classifier and Generator
    - Explanation Factor
    - Minimum Risk Training
- Experiment results

Overview

- Motivation
- Proposed solution
- Generative Explanation Framework (GEF)
    - **Intuition**
    - Base Classifier and Generator
    - Explanation Factor
    - Minimum Risk Training
- Experiment results

Generative Explanation Framework

Intuition - GEF

- Build a classifier model to do text classification

Generative Explanation Framework

Intuition - GEF

- Build a classifier model to do text classification
- Obtain good fine-grained information from raw text to explain the prediction (and improve performance?)

Generative Explanation Framework

Intuition - GEF

- **Q:** What is a fine-grained explanation?

Generative Explanation Framework

Intuition - GEF

- **Q:** What is a fine-grained explanation?

Generative Explanation Framework

Intuition - GEF

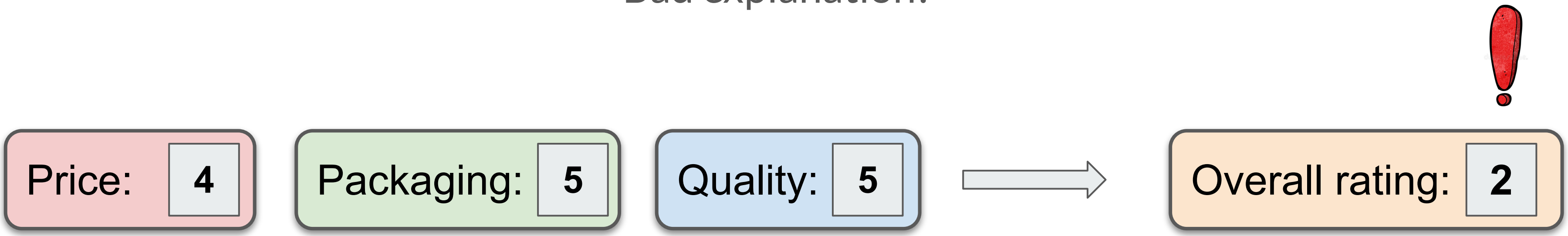
- Q: What is a **good** fine-grained explanation?

Generative Explanation Framework

Intuition - GEF

- **Q:** What is a **good** fine-grained explanation?

Bad explanation!

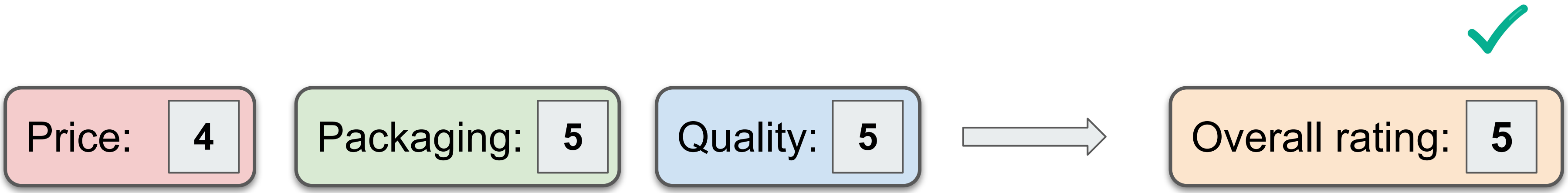


Generative Explanation Framework

Intuition - GEF

- **Q:** What is a **good** fine-grained explanation?

Good explanation!



Generative Explanation Framework

Data

- **Example:** Rating a product
  - Review text: “The phone feels sturdy, looks premium,.......”
  - Fine-grained/Golden explanations:

    Price: 4 | Packaging: 5 | Quality: 5

  - Output:

    Overall rating: 5

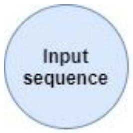
Overview

- Motivation
- Proposed solution
- **Generative Explanation Framework (GEF)**
    - Intuition
    - **Base Classifier and Generator**
    - Explanation Factor
    - Minimum Risk Training
- Experiment results

Generative Explanation Framework

Base Classifier and Generator

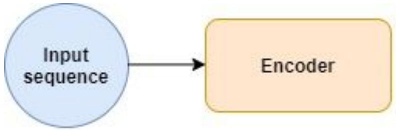
- Base classifier
    - Encoder-Predictor architecture



Generative Explanation Framework

Base Classifier and Generator

- Base classifier
  - Encoder-Predictor architecture



Generative Explanation Framework

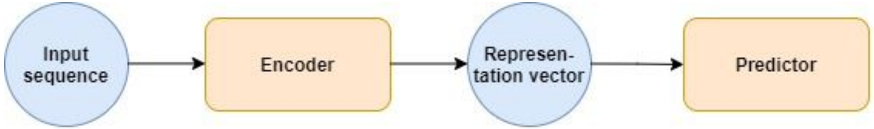
Base Classifier and Generator

- Base classifier
  - Encoder-Predictor architecture

Generative Explanation Framework

Base Classifier and Generator

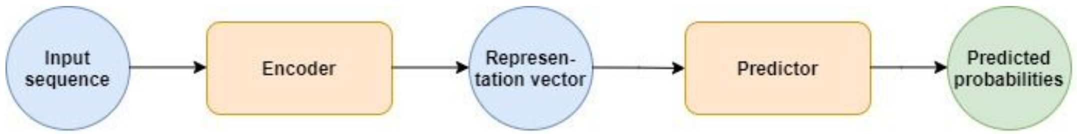
- Base classifier
  - Encoder-Predictor architecture



Generative Explanation Framework

Base Classifier and Generator

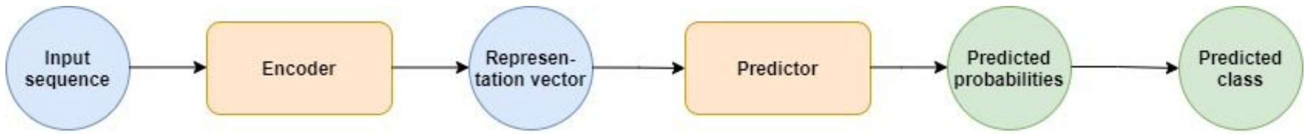
- Base classifier
  - Encoder-Predictor architecture



Generative Explanation Framework

Base Classifier and Generator

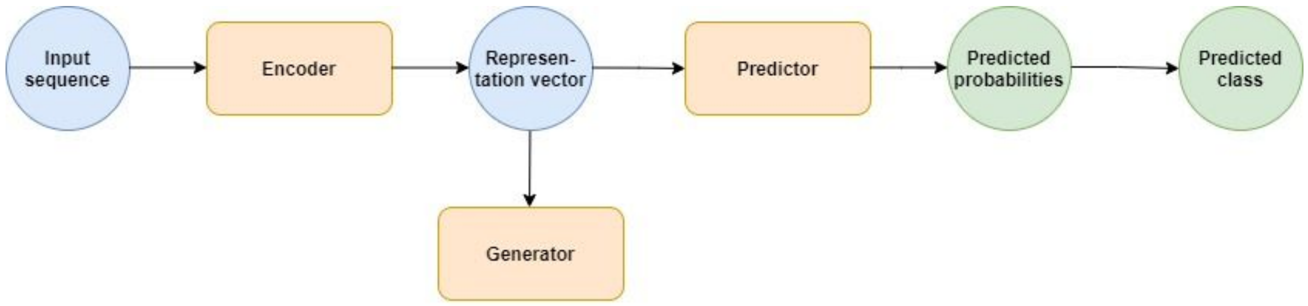
- Base classifier
  - Encoder-Predictor architecture



Generative Explanation Framework

Base Classifier and Generator

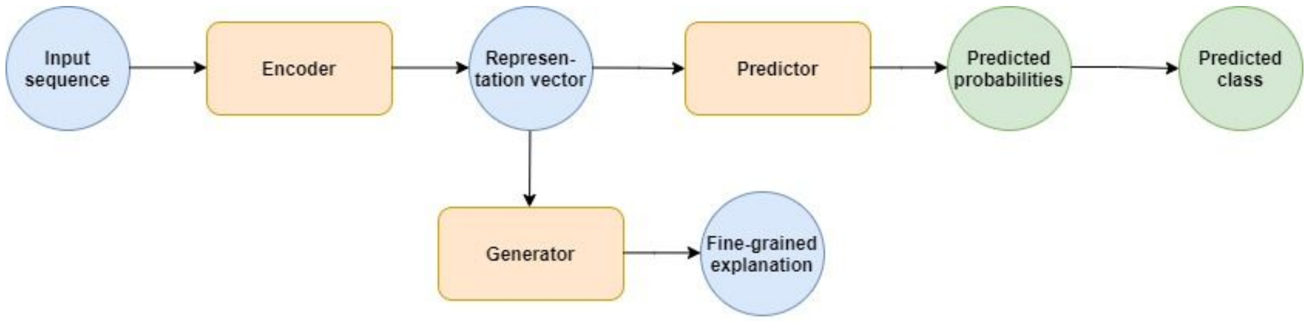
- Generator



Generative Explanation Framework

Base Classifier and Generator

- Generator



Generative Explanation Framework

Notations

Input sequence of texts: $S\{s_1, s_2, \ldots, s_{|S|}\}$

Output category: $y_i (i \in [1, 2, \ldots, N])$

Generative Explanation Framework

Algorithm so far...

- **Step -1:** Encode the input sequences to get representation vectors

$$v_e = Encoder([s_1, s_2, \cdots, s_{|S|}])$$

Generative Explanation Framework

Algorithm so far...

- **Step -1:** Encode the input sequences to get representation vectors

$$v_e = Encoder([s_1, s_2, \cdots, s_{|S|}])$$

- **Step -2:** A predictor takes those vectors as input and outputs probabilities by using softmax

$$P_{pred} = Predictor(v_e)$$

$$y = \arg\max_i (P_{pred,i})$$

Generative Explanation Framework

Algorithm so far...

- **Step -3:** Feed the representation vectors to a Generator G to generate fine-grained explanations

$$e_c = f_G(W_G \cdot v_e + b_G)$$

Generative Explanation Framework

Algorithm so far...

- **Step -3:** Feed the representation vectors to a Generator G to generate fine-grained explanations

$$e_c = f_G(W_G \cdot v_e + b_G)$$

- **Loss:**

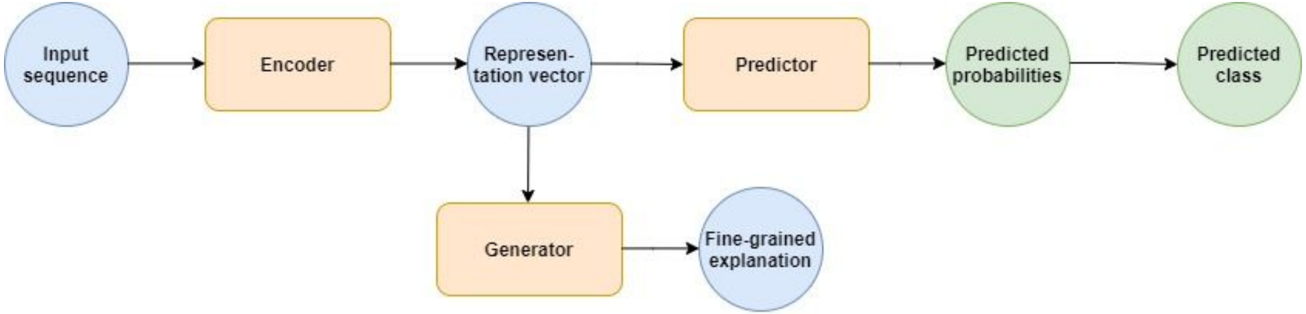
  Overall loss = Classification loss + Explanation generation loss

$$\mathcal{L}(e_g, S, \theta) = \mathcal{L}_p + \mathcal{L}_e$$

Generative Explanation Framework

Base Classifier and Generator

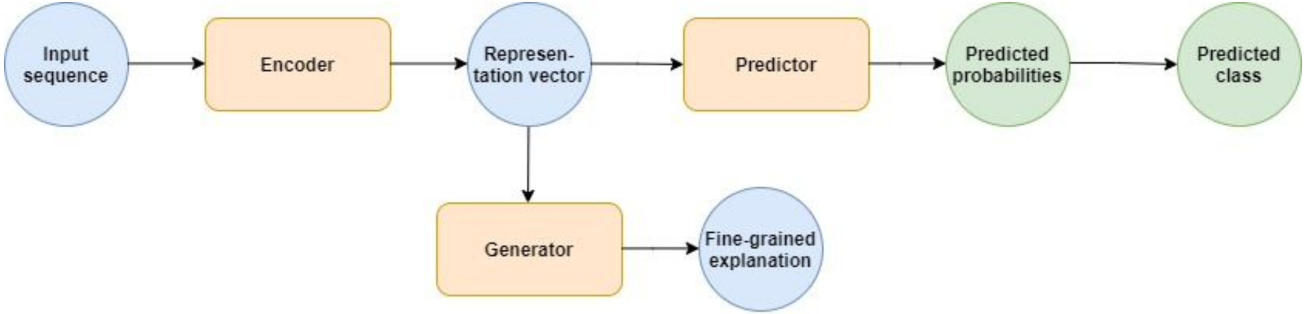
- Encoder-Predictor and Generator



Generative Explanation Framework

Base Classifier and Generator

- Encoder-Predictor and Generator

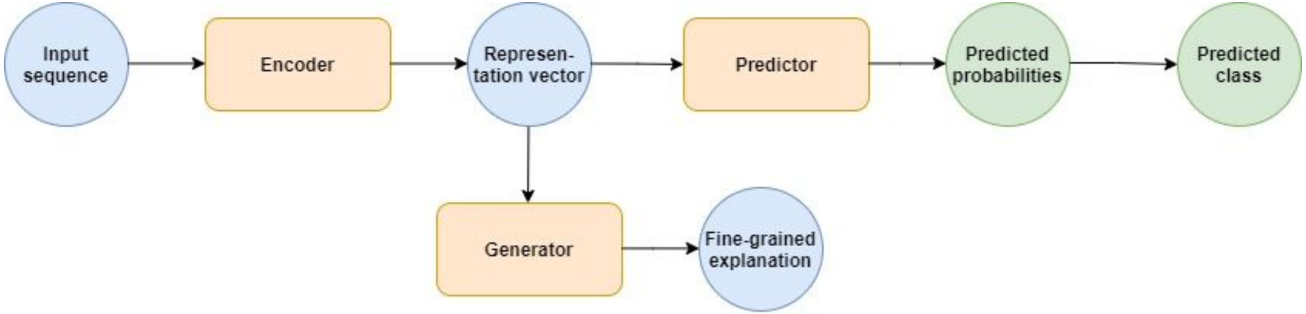


Generative explanations seem to be independent of the predicted overall results

Generative Explanation Framework

Base Classifier and Generator

- Encoder-Predictor and Generator



Generative explanations seem to be independent of the predicted overall results

Use Explanation Factor!

Overview

- Motivation
- Proposed solution
- **Generative Explanation Framework (GEF)**
  - Intuition
  - Base Classifier and Generator
  - **Explanation Factor**
  - Minimum Risk Training
- Experiment results

Generative Explanation Framework

Explanation Factor

Generative explanations seem to be independent of the predicted overall results

Generative Explanation Framework

Explanation Factor

Generative explanations seem to be independent of the predicted overall results

- **Example:**

  “The product is good to use” ⇒ Overall rating = 5? or =4??

Generative Explanation Framework

Explanation Factor

Generative explanations seem to be independent of the predicted overall results

- **Example:**
  "The product is good to use" ⇒ Overall rating = 5? or =4??

Generative Explanation Framework

Explanation Factor

- **Idea:** Pre-train a classifier which learns to predict by directly taking the explanations as input

Generative Explanation Framework

Explanation Factor

- **Idea:** Pre-train a classifier which learns to predict by directly taking the explanations as input

Generative Explanation Framework

Explanation Factor

- **Idea:** Pre-train a classifier which learns to predict by directly taking the explanations as input

- This should predict the overall results more accurately than the base model that takes raw text as the input

Generative Explanation Framework

Explanation Factor

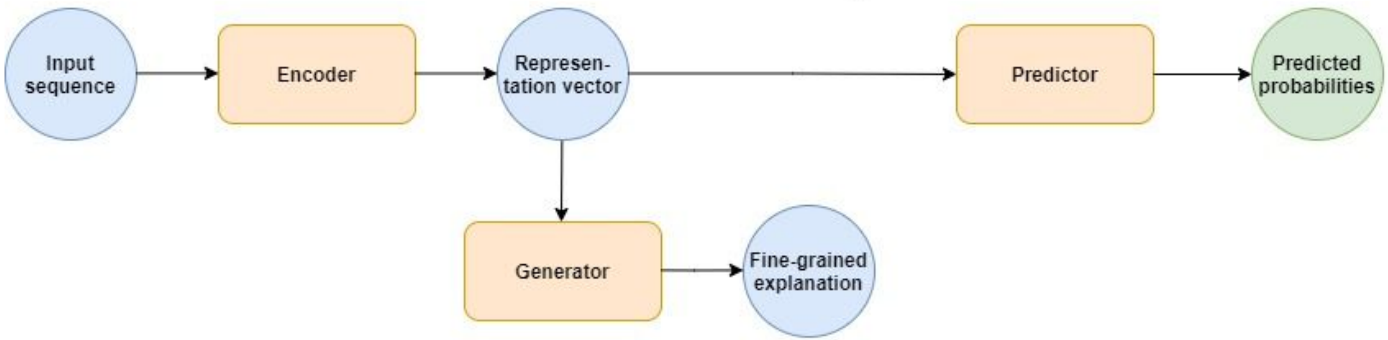
- **Idea:** Pre-train a classifier which learns to predict by directly taking the explanations as input

- The classifier also helps in providing a strong guidance for the text encoder to generate a more informative representation vector

Generative Explanation Framework

Explanation Factor

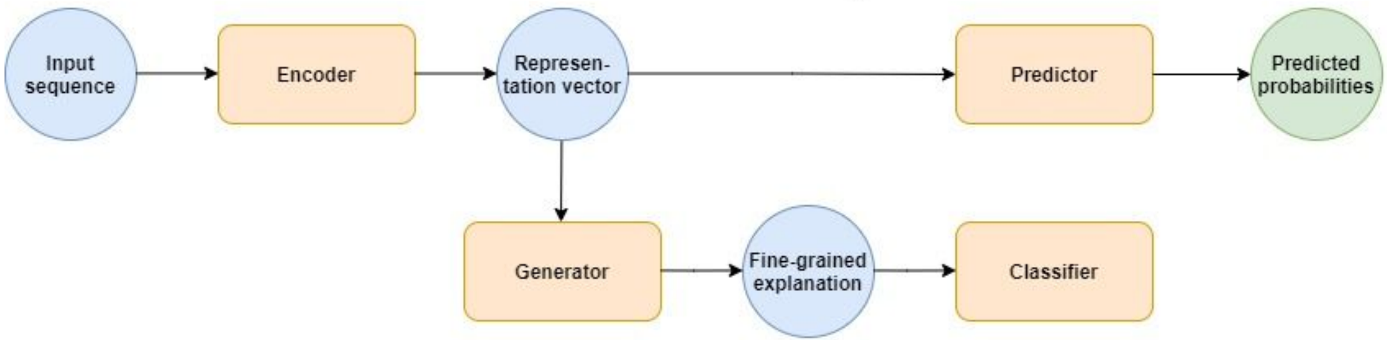
- Base classifier and Generator



Generative Explanation Framework

Explanation Factor

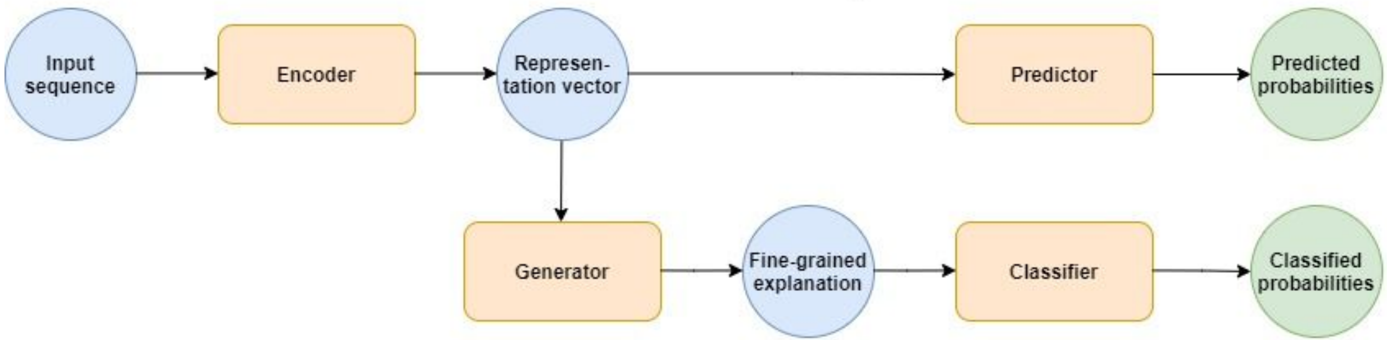
- (Pre-trained) Classifier



Generative Explanation Framework

Explanation Factor

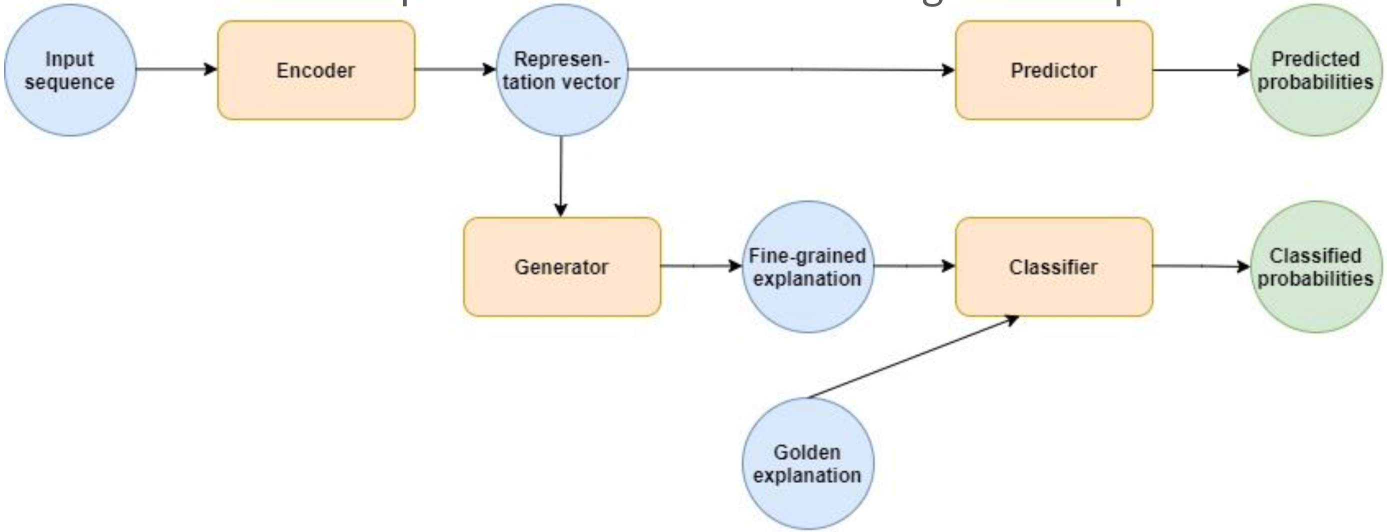
- (Pre-trained) Classifier



Generative Explanation Framework

Explanation Factor

- (Pre-trained) Classifier
  - Golden explanation ⇔ 'Actual' fine-grained explanation*

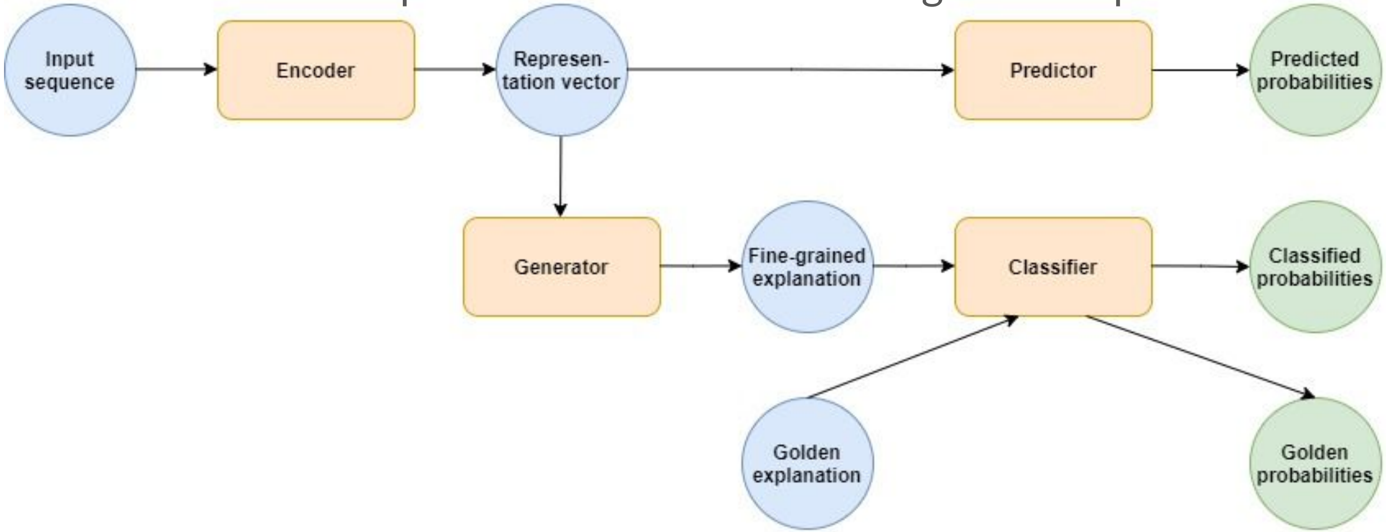


* the fine-grained information collected/scraped by the authors in this paper

Generative Explanation Framework

Explanation Factor

- (Pre-trained) Classifier
  - Golden explanation ⇔ 'Actual' fine-grained explanation*



* the fine-grained information collected/scraped by the authors in this paper

Generative Explanation Framework

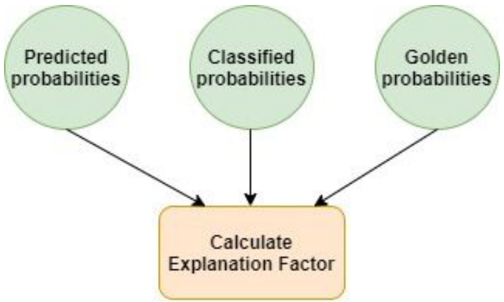
Explanation Factor

- Calculate **Explanation factor** from the obtained probabilities

Generative Explanation Framework

Explanation Factor

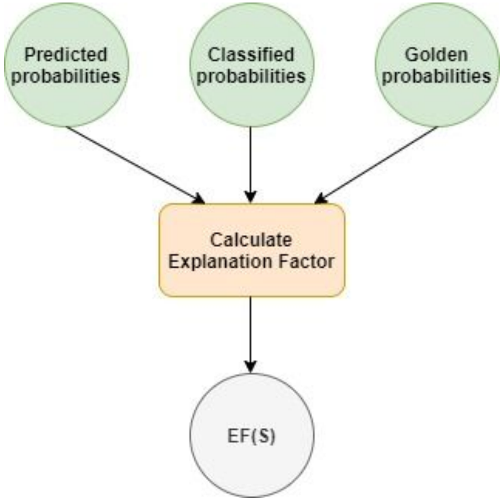
- Calculate **Explanation factor** from the obtained probabilities



Generative Explanation Framework

Explanation Factor

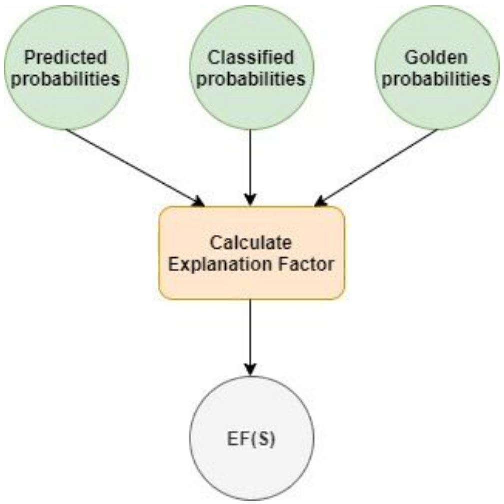
- Calculate **Explanation factor** from the obtained probabilities



Generative Explanation Framework

Explanation Factor

- Calculate **Explanation factor** from the obtained probabilities



But how?

Generative Explanation Framework

Explanation Factor

1. Obtained probabilities:

$$P_{pred} = Predictor(v_e)$$

$$P_{classified} = softmax(f_C(W_C \cdot e_c + b_C))$$

$$P_{gold} = softmax(f_C(W_C \cdot e_g + b_C))$$

Generative Explanation Framework

Explanation Factor

1. Obtained probabilities:

$$P_{pred} = Predictor(v_e)$$
$$P_{classified} = softmax(f_C(W_C \cdot e_c + b_C))$$
$$P_{gold} = softmax(f_C(W_C \cdot e_g + b_C))$$

2. Extract ground-truth probability $\tilde{P}_{classified}, \tilde{P}_{pred}, \tilde{P}_{gold}$ from the obtained probabilities

Generative Explanation Framework

Explanation Factor

1. Obtained probabilities:

$$P_{pred} = Predictor(v_e)$$

$$P_{classified} = softmax(f_C(W_C \cdot e_c + b_C))$$

$$P_{gold} = softmax(f_C(W_C \cdot e_g + b_C))$$

2. Extract ground-truth probability $\tilde{p}_{classified}, \tilde{p}_{pred}, \tilde{p}_{gold}$ from the obtained probabilities

Eg: $P_{classified}$ = [0.5 0.2 0.1 0.2] ; when $y = 2$ ⇒ $\tilde{p}_{classified}$ = 0.2

Generative Explanation Framework

Explanation Factor

3. Explanation factor:

$$EF(S) = |\tilde{p}_{classified} - \tilde{p}_{gold}| + |\tilde{p}_{classified} - \tilde{p}_{pred}|$$

Generative Explanation Framework

Explanation Factor

3. Explanation factor:

$$EF(S) = |\tilde{p}_{classified} - \tilde{p}_{gold}| + |\tilde{p}_{classified} - \tilde{p}_{pred}|$$

- $|\tilde{p}_{classified} - \tilde{p}_{gold}|$ $\rightarrow$ distance between the generated explanations and the golden explanations

Generative Explanation Framework

Explanation Factor

3. Explanation factor:

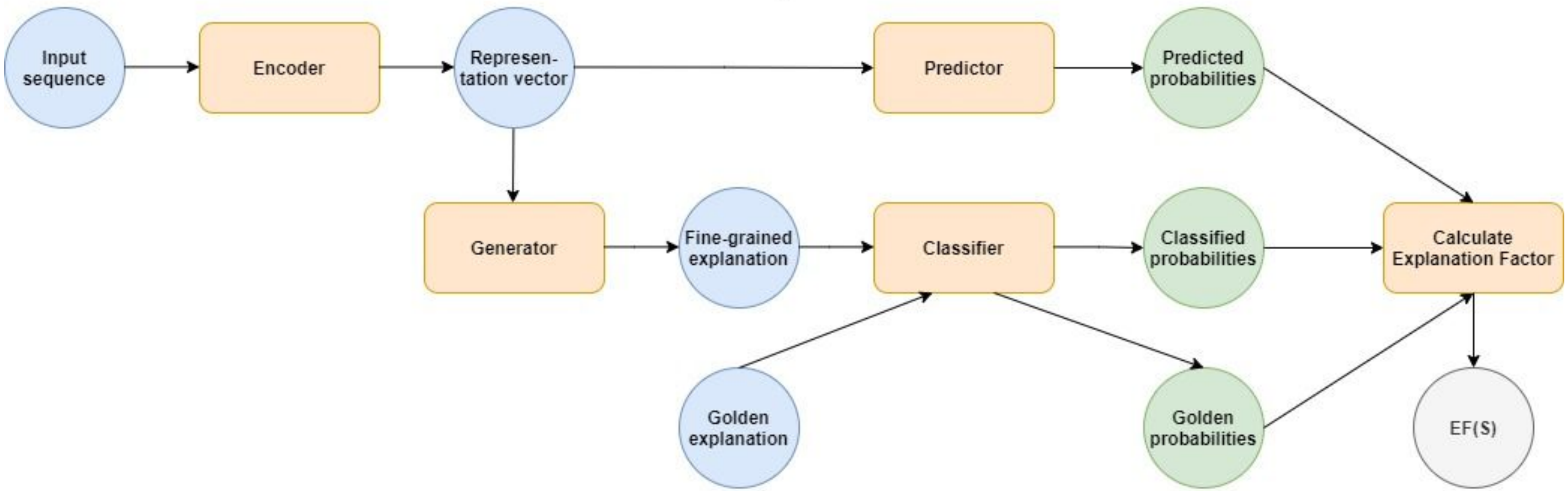
$$EF(S) = |\tilde{p}_{classified} - \tilde{p}_{gold}| + |\tilde{p}_{classified} - \tilde{p}_{pred}|$$

- $|\tilde{p}_{classified} - \tilde{p}_{gold}|$ $\rightarrow$ distance between the generated explanations and the golden explanations
- $|\tilde{p}_{classified} - \tilde{p}_{pred}|$ $\rightarrow$ relevance between the generated explanations and the original text

Generative Explanation Framework

Explanation Factor

- The GEF framework



Generative Explanation Framework

Algorithm so far...

- **Step -1 to 3:** Encoder, Predictor, Generator

Generative Explanation Framework

Algorithm so far...

- **Step -1 to 3:** Encoder, Predictor, Generator

- **Step -4:** Feed the generated fine-grained explanations to the pre-trained* classifier to obtain $P_{classified}$

*the pre-trained classifier is trained to output the right class as output using the golden explanations as input

Generative Explanation Framework

Algorithm so far...

- **Steps -1 to 3:** Encoder, Predictor, Generator

- **Step -4:** Feed the generated fine-grained explanations to the pre-trained* classifier to obtain $P_{classified}$

- **Step -5:** Feed the golden explanations to the pre-trained* classifier to obtain P_{gold}

*the pre-trained classifier is trained to output the right class as output using the golden explanations as input

Generative Explanation Framework

Algorithm so far...

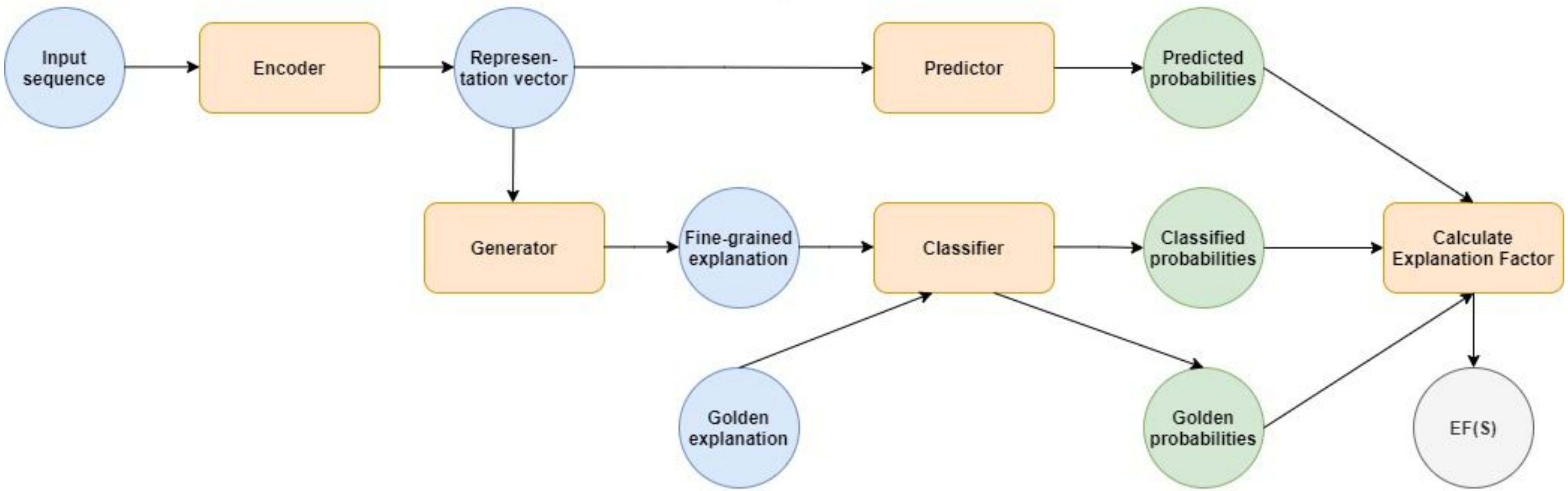
- **Step -6:** Calculate explanation factor from the obtained probabilities

$$EF(S) = |\tilde{p}_{classified} - \tilde{p}_{gold}| + |\tilde{p}_{classified} - \tilde{p}_{pred}|$$

Generative Explanation Framework

Explanation Factor

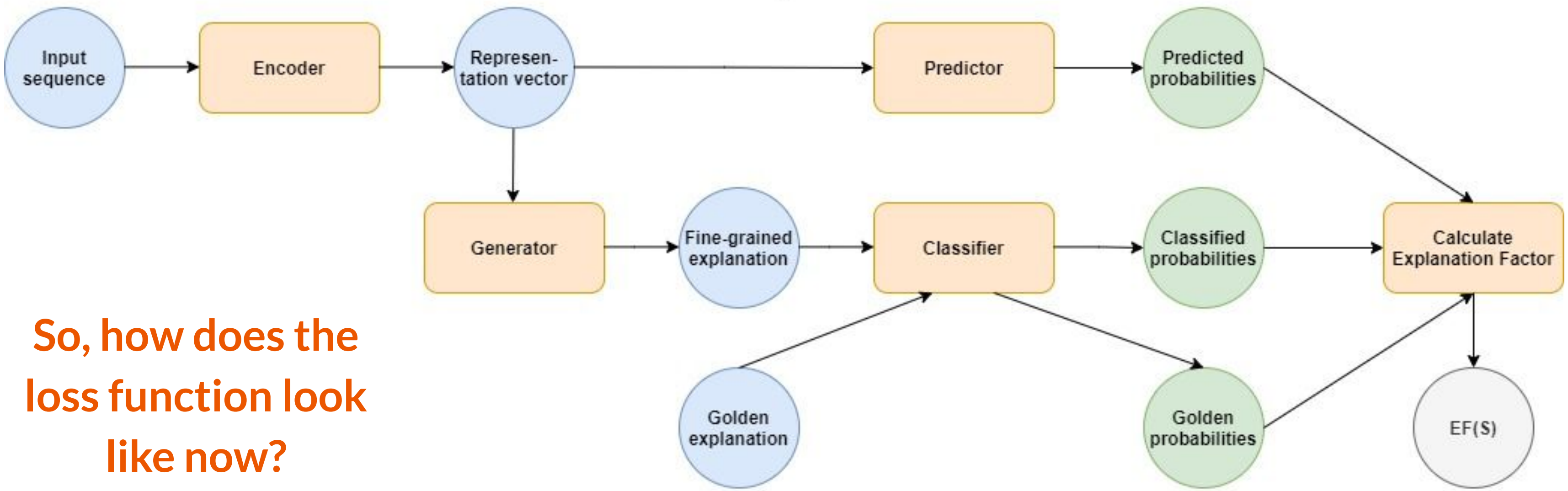
- The GEF framework



Generative Explanation Framework

Explanation Factor

- The GEF framework



So, how does the loss function look like now?

Overview

- Motivation
- Proposed solution
- **Generative Explanation Framework (GEF)**
  - Intuition
  - Base Classifier and Generator
  - Explanation Factor
  - **Minimum Risk Training**
- Experiment results

Generative Explanation Framework

Minimum Risk Training

- Optimization: Minimum Risk Training (**MRT**)

Generative Explanation Framework

Minimum Risk Training

- Optimization: Minimum Risk Training (**MRT**)
  - Minimize the expected loss i.e., risk over the training data

Generative Explanation Framework

Minimum Risk Training

- Optimization: Minimum Risk Training (**MRT**)
  - Minimize the expected loss i.e., risk over the training data

$$\mathcal{L}_{MRT}(e_g, S, \theta) = \sum_{(e_g, S) \in D} \mathcal{L}(e_g, S, \theta) EF(S)$$

Generative Explanation Framework

Minimum Risk Training

- Optimization: Minimum Risk Training (**MRT**)

$$\mathcal{L}_{MRT}(e_g, S, \theta) = \sum_{(e_g, S) \in D} \mathcal{L}(e_g, S, \theta) EF(S)$$

Generative Explanation Framework

Minimum Risk Training

- Optimization: Minimum Risk Training (**MRT**)

$$\mathcal{L}_{MRT}(e_g, S, \theta) = \sum_{(e_g, S) \in D} \mathcal{L}(e_g, S, \theta) EF(S)$$

  - $\mathcal{L}(e_g, S, \theta)$ - Classification loss + Explanation generation loss

Generative Explanation Framework

Minimum Risk Training

- Optimization: Minimum Risk Training (**MRT**)

$$\mathcal{L}_{MRT}(e_g, S, \theta) = \sum_{(e_g, S) \in D} \mathcal{L}(e_g, S, \theta) EF(S)$$

  - $\mathcal{L}(e_g, S, \theta)$ - Classification loss + Explanation generation loss
  - EF(S) is treated as the semantic distance of input texts, generated explanations and golden explanations

Generative Explanation Framework

Minimum Risk Training

- Optimization: Minimum Risk Training (**MRT**)

$$\mathcal{L}_{MRT}(e_g, S, \theta) = \sum_{(e_g, S) \in D} \mathcal{L}(e_g, S, \theta) EF(S)$$

  - $\mathcal{L}_{MRT}$ can be zero or close to zero when $\tilde{p}_{classified}, \tilde{p}_{pred}, \tilde{p}_{gold}$ are close

Generative Explanation Framework

Minimum Risk Training

- Optimization: Minimum Risk Training (**MRT**)

$$\mathcal{L}_{MRT}(e_g, S, \theta) = \sum_{(e_g,S)\in D} \mathcal{L}(e_g, S, \theta) EF(S)$$

  - $\mathcal{L}_{MRT}$ can be zero or close to zero when $\tilde{p}_{classified}, \tilde{p}_{pred}, \tilde{p}_{gold}$ are close
  - But, this cannot guarantee that generated explanations are close to the golden explanations

Generative Explanation Framework

Minimum Risk Training

- Optimization: Minimum Risk Training (**MRT**)
    - To avoid total degradation of loss, the final loss function will be:

Generative Explanation Framework

Minimum Risk Training

- Optimization: Minimum Risk Training (**MRT**)
  - To avoid total degradation of loss, the final loss function will be:

$$\mathcal{L}_{final} = \sum_{(e_g,S)\in D} \mathcal{L} + \mathcal{L}_{MRT}$$

Generative Explanation Framework

Minimum Risk Training

- Optimization: Minimum Risk Training (**MRT**)
  - To avoid total degradation of loss, the final loss function will be:

$$\mathcal{L}_{final} = \sum_{(e_g, S) \in D} \mathcal{L} + \mathcal{L}_{MRT}$$

  - Final loss function = Explanation generation loss + MRT loss

Generative Explanation Framework

Minimum Risk Training

- Optimization: Minimum Risk Training (**MRT**)
  - To avoid total degradation of loss, the final loss function will be:

$$\boxed{\mathcal{L}_{final} = \sum_{(e_g,S)\in D} \mathcal{L} + \mathcal{L}_{MRT}}$$

  - Final loss function = Explanation generation loss + MRT loss

Generative Explanation Framework

THE algorithm

- **Steps - 1 to 3:** Encoder, Predictor, Generator
- **Steps - 4 to 6:** Obtain the probabilities and calculate Explanation Factor
- **Step - 7:** Calculate the final loss and optimize the model

Overview

- Motivation
- Proposed solution
- Generative Explanation Framework (GEF)
  - Intuition
  - Base Classifier and Generator
  - Explanation Factor
  - Minimum Risk Training
- **Experiment results**

Experiment results

- Fine-grained explanations are in different forms
  - Text
  - Numeric

Experiment results

- Fine-grained explanations are in different forms
  - Text - PCMag review dataset
  - Numeric - Skytrax airline review

Experiment results

- Fine-grained explanations are in different forms
  - Text - PCMag review dataset
  - Numeric - Skytrax airline review
- GEF has been applied to both forms of explanations using different base models

Experiment results

- Fine-grained explanations are in different forms
  - Text - PCMag review dataset
  - Numeric - Skytrax airline review
- GEF has been applied to both forms of explanations using different base models
- Experimental settings are set the same for base model and base model+GEF for easy comparison of the performance

Experiment results

Experimental settings

- **Tokenizer** : Stanford Tokenizer
- **Embedding** : GloVe
- **Optimizer** : Adam
- **Stopping criteria** : Stop updating when the classification loss reaches a certain threshold [since generation loss > classification loss for text explanations]

Experiment results

Text explanations

- PCMag dataset:
  - Long review text for electronic products
  - Three short comments
    - positive, negative, neutral
  - Overall rating score
    - {1.0, 1.5, 2.0, ..., 5.0}
- Filter: review text with >70 sentences or comments >75 tokens have been removed
- Train/Dev/Test: 10919/1373/1356

Experiment results

Text explanations

- **Base model**: CVAE [Conditional Variational AutoEncoder]
- **Proposed model**: CVAE + GEF
- **Classifier:** Skip-connected model with bidirectional GRU-RNN layers
- **Evaluation metric (for text generation):** BLEU score
- **Evaluation metric (for classification):** top-1, top-3 accuracy

Experiment results

Text explanations

		BLEU-1	BLEU-2	BLEU-3	BLEU-4
Pos.	CVAE	36.1	13.5	3.7	2.2
	CVAE+GEF	**40.1**	**15.6**	**4.5**	**2.6**
Neg.	CVAE	33.3	14.1	3.1	2.2
	CVAE+GEF	**35.9**	**16.0**	**4.0**	**2.9**
Neu.	CVAE	30.0	8.8	2.0	1.2
	CVAE+GEF	**33.2**	**10.2**	**2.5**	**1.5**

Table 4: BLEU scores for generated explanations.

- **Base model**: CVAE [Conditional Variational AutoEncoder]
- **Proposed model**: CVAE + GEF
- **Classifier:** Skip-connected model with bidirectional GRU-RNN layers
- **Evaluation metric** **(for text generation):** BLEU score
- **Evaluation metric** **(for classification):** top-1, top-3 accuracy

Experiment results

Text explanations

	Acc% (Dev)	Acc% (Test)
CVAE	42.07	42.58
CVAE+GEF	**44.04**	**43.67**
Oracle	46.43	46.73

Table 5: Classification accuracy on PCMag Review Dataset. *Oracle* means if we feed ground-truth text explanations to the Classifier C, the accuracy C can achieve to do classification. *Oracle* confirms our assumption that explanations can do better in classification than the original text.

- **Base model**: CVAE [Conditional Variational AutoEncoder]
- **Proposed model**: CVAE + GEF
- **Classifier:** Skip-connected model with bidirectional GRU-RNN layers
- **Evaluation metric** **(for text generation)**: BLEU score
- **Evaluation metric** **(for classification)**: top-1, top-3 accuracy

Experiment results

Numerical explanations

- Skytrax dataset:
  - Review text of airlines
  - Five sub-field scores - [0-5]
    - Seat comfortability, cabin stuff, food, in-flight environment, ticket value
  - Overall rating score - [1-10]
- Filter: review text with >300 tokens
- Train/Dev/Test: 21676/2710/2709

Experiment results

Numerical explanations

- **Base model:** LSTM and CNN
- **Proposed model:** LSTM + GEF
  CNN + GEF
- **Evaluation metric:** Accuracy

Experiment results

Numerical explanations

	s%	c%	f%	i%	t%
LSTM	46.59	52.27	43.74	41.82	45.04
LSTM+GEF	**49.13**	**53.16**	**46.29**	**42.34**	**48.25**
CNN	46.22	51.83	44.59	43.34	46.88
CNN+GEF	**49.80**	**52.49**	**48.03**	**44.67**	**48.76**

Table 6: Accuracy of sub-field numerical explanations on Skytrax User Reviews Dataset. s, c, f, t, v stand for seat comfortability, cabin stuff, food, in-flight environment and ticket value, respectively.

- **Base model:** LSTM and CNN
- **Proposed model:** LSTM + GEF

  CNN + GEF
- **Evaluation metric**: Accuracy

Experiment results

Numerical explanations

	Acc%	Top-3 Acc%
LSTM	38.06	76.89
LSTM+GEF	**39.20**	**77.96**
CNN	37.06	76.85
CNN+GEF	**39.02**	**79.07**
Oracle	45.00	83.13

Table 7: Classification accuracy on Skytrax User Reviews Dataset. *Oracle* means if we feed ground-truth numerical explanation to the Classifier C, the accuracy C can achieve to do classification.

- **Base model:** LSTM and CNN
- **Proposed model:** LSTM + GEF

  CNN + GEF
- **Evaluation metric:** Accuracy

Experiment results

Human evaluation

	Win%	Lose%	Tie%
CVAE+GEF	51.37	42.38	6.25

Table 8: Results of human evaluation. Tests are conducted between the text explanations generated by basic CVAE and CVAE+GEF.

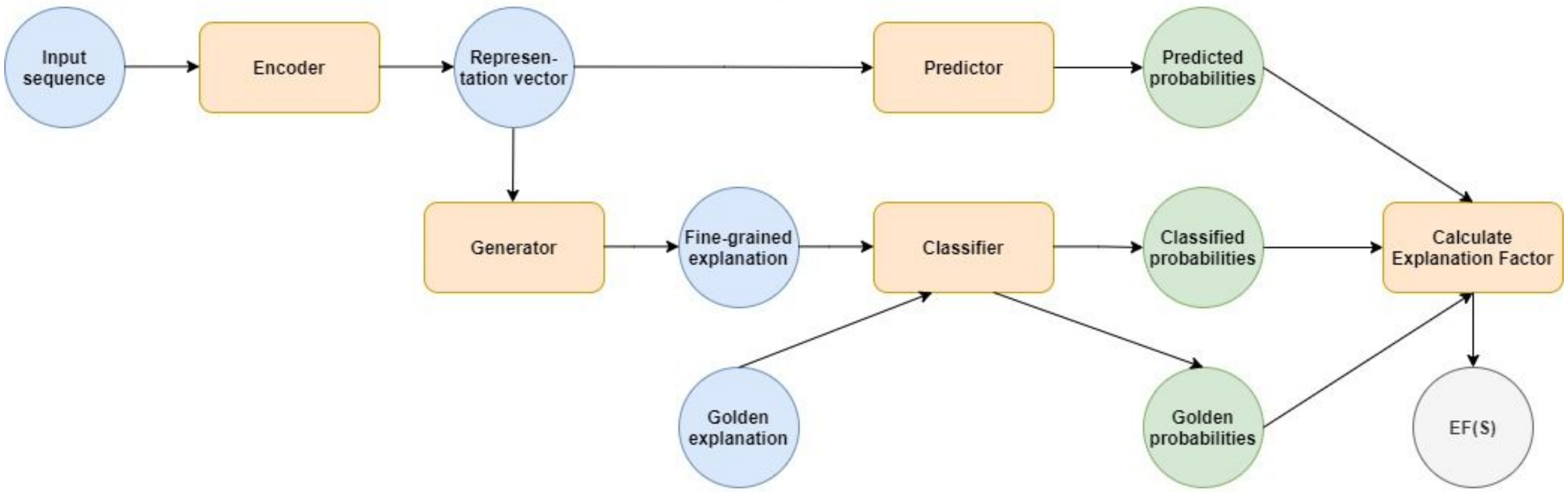
- Crowdsourced judges in Amazon Mechanical Turk
- **# samples** : 100 items
- **# judges** : 5
- All items are correctly classified using both basic model and using GEF

Summary

- Rationale is essential for an NLP system to be reliable
- A novel Generative Explanation Framework is introduced
- GEF uses the generated fine-grained information to leverage the performance of the text classification model

Summary

- The GEF framework



The End